Privacy-Preserving Linkage of Genomic and Clinical Data Sets

Baker, Dixie B.; Knoppers, Bartha M.; Phillips, Mark; van Enckevort, David; Kaufmann, Petra; Lochmuller, Hanns; Taruscio, Domenica

Published in:
IEEE/ACM Transactions on Computational Biology and Bioinformatics

DOI:
10.1109/TCBB.2018.2855125

Publication date:
2019

Privacy-Preserving Linkage of Genomic and Clinical Data Sets

Dixie B. Baker, Bartha M. Knoppers, Mark Phillips, David van Enckevort, Petra Kaufmann, Hanns Lochmuller, and Domenica Taruscio

Abstract—The capacity to link records associated with the same individual across data sets is a key challenge for data-driven research. The challenge is exacerbated by the potential inclusion of both genomic and clinical data in data sets that may span multiple legal jurisdictions, and by the need to enable re-identification in limited circumstances. Privacy-Preserving Record Linkage (PPRL) methods address these challenges. In 2016, the Interdisciplinary Committee of the International Rare Diseases Research Consortium (IRDiRC) launched a task team to explore approaches to PPRL. The task team is a collaboration with the Global Alliance for Genomics and Health (GA4GH) Regulatory and Ethics and Data Security Work Streams, and aims to prepare policy and technology standards to enable highly reliable linking of records associated with the same individual without disclosing their identity except under conditions in which the use of the data has led to information of importance to the individual's safety or health, and applicable law allows or requires the return of results. The PPRL Task Force has examined the ethico-legal requirements, constraints, and implications of PPRL, and has applied this knowledge to the exploration of technology methods and approaches to PPRL. This paper reports and justifies the findings and recommendations thus far.

Index Terms—Data matching, privacy, privacy-preserving record linkage, record linkage

1 Introduction

Privacy-Preserving Record Linkage (PPRL) [1] addresses two primary challenges that lie at the intersection of biomedical research and clinical practice:

1. The de-duplication and linking of datasets for use by researchers, without disclosing the participant's identity; and
2. The re-identification of research participants for clinical purposes, such as to return results that may be useful in clinical diagnosis or treatment.

In 2016, the Global Alliance for Genomics and Health (GA4GH) (genomicsandhealth.org) launched a task team to explore ethical questions, regulatory requirements, and technological methods and approaches related to PPRL. The task team is a collaboration in which the GA4GH (Regulatory and Ethics Work Stream and the Data Security Work Stream) is preparing policy and technology standards, together with the Interdisciplinary Committee of the International Rare Diseases Research Consortium (IRDiRC) to enable highly reliable linking of coded data records associated with the same individual without disclosing the identity of that individual except under conditions in which the use of the data has led to information of importance to the individual's safety or health, and applicable law allows or requires the return of results.

The primary motivation of the GA4GH in this endeavour is its conviction that because linkage enables the creation, availability, and precision of data, it therefore improves the quality of both research and the health care provided to people. The GA4GH believes that this reinforces the right to share in scientific advancement and its benefits as guaranteed by Article 27 of the Universal Declaration of Human Rihts, as mobilized in the GA4GH *Framework for Responsible Sharing of Genomic and Health-Related Data* [2].

- *D. B. Baker is with Martin, Blanck, and Associates, Arlington, VA 22314. E-mail: dixie.baker@martin-blanck.com.*
- *B. M. Knoppers and M. Phillips are with the Centre of Genomics and Policy, McGill University, Quebec H3A 0G1, Canada. E-mail: {bartha.knoppers, mark.phillips2}@mcgill.ca.*
- *D. van Enckevort is with the University Medical Center Groningen, Groning-en, Netherlands. E-mail: david.van.enckevort@umcg.nl.*
- *P. Kaufmann is with the National Institutes of Health, Bethesda 20892, MD. E-mail: petra.kaufmann@nih.gov.*
- *H. Lochmuller is with Newcastle University, Newcastle Upon Tyne, United Kingdom. E-mail: hanns.lochmuller@newcastle.ac.uk.*
- *D. Taruscio is with the Istituto Superiore di Sanità, Rome, Italy. E-mail: domenica.taruscio@iss.it.*

Manuscript received 11 June 2018; accepted 21 June 2018. Date of publication 31 July 2018; date of current version 5 Aug. 2019.
(Corresponding author: Mark Phillips.)
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TCBB.2018.2855125

2 Ethical and Legal Considerations

2.1 Sensitivity Considerations or Centralization

The ability to conveniently link data can itself dramatically increase the risk of breach, because it carries with it a corresponding boost in various adversaries' motivation to access the more valuable data sets. State and non-state actors alike will hardly be able to pass up the chance to access comprehensive, centralized (or centralizable) global repositories. Any PPRL system's designers should first carefully analyze the risks, benefits, and available safeguards based on a variety of threats.

Legal frameworks recognize risk as inherent to linkage. Personal data held for research purposes in the Canadian province of British Columbia, for example, can be linked only

if "any data linkage is not harmful to the individuals who are the subjects, . . . and benefits . . . are clearly in the public interest" [3], [4], [5]. Article 35 of the European Union's General Data Protection Regulation (GDPR) has the effect of requiring that a data protection impact assessment be carried out prior to putting in place an international PPRL system" [7].

Data aggregation through linkage runs the risk of inadvertently transforming into identifiable data, data that were not previously perceived as reasonably foreseeably identifiable. Linkage methods should thus either include metrics for measuring the level of protection or offer alternative safeguards.

2.2 Generating Linkage Data

The simplest method of linking data about an individual is to assign to the individual a unique identifier that is derived from a relatively immutable set of the individual's personal data and that would irrefutably be associated with the individual and her data wherever they may go. Although several countries have moved toward the approach, it has often been accompanied by controversy, and this approach is neither legally nor politically feasible in global health-related data sharing. The most direct impediment is that a number of laws require or recommend that identifiers not be generated on the basis of personal information [6], [7].

A more privacy-conscious approach to linkage is to associate the immutable personal data with a randomly generated pseudonym that will serve as the unique identifier associated with all the individual's records. In this way, linking the pseudonym with anything about the person will be impossible without first having access to the relevant immutable data.

But this approach suffers from its own legal and practical shortcomings. Since the goal of the exercise is to link as many data sets as possible, and since data collection is rapidly increasing, a correspondingly larger number of people would necessarily have access to the immutable data in question, thus creating a major vulnerability in the privacy-protection scheme. The law in many jurisdictions reflects the reality of this danger. For example, since 1999, the US Congress has in its appropriations bills consistently prohibited the use of federal funds to create a standard, unique health identifier.

In the European Union, although the *General Data Protection Regulation*, like the *Data Protection Directive* before it, empowers member states to adopt frameworks for an "identifier of general application," few have done so. The provision in the *Regulation* adds a new condition that despite any national framework, such identifiers "shall be used only under appropriate safeguards for the rights and freedoms of the data subject pursuant to this Regulation." The existing interpretive guidance additionally suggests avoiding the use of the same pseudonym across different datasets [7].

PPRL will have to turn to methods other than those that suffer from the shortcomings described above.

2.3 Participant Withdrawal

The design of an effective PPRL system should also include robust support for participant withdrawal. This right arises strongly in research ethics, medical ethics, and data protection. The GDPR explicitly requires that it must "be as easy to withdraw as to give the consent" [7]. The right to informed consent, whether in research ethics, medical liability, or data protection, always includes an inalienable right to revoke consent at any moment, limited only by the impossibility of changing the past.

Withdrawal poses special challenges with respect to the use of immutable data as identifiers, hash values derived from immutable data, and Bloom filters [1] that use immutable identifiers to link data. The right to withdraw also poses pernicious difficulties in distributed, complex systems where distinct entities link, aggregate, or share existing data sets without a uniformly enforced governance policy that specifically enables participant withdrawal.

Although retroactive withdrawal is generally not possible, prospective withdrawal is essential. Systems should be designed to allow participants or patients to withdraw their consent to the processing of their personal data without the risk of their previously linked data "re-emerging" when their data are re-entered into the system in some future time. This can be a challenge for hash-based systems in some configurations, particularly when the underlying metadata are simply removed from the index and not removed from the system. These difficulties are not technically insurmountable, but a PPRL system should include a means of ensuring that the participant is able to withdraw her consent at any time.

2.4 Returning Results to Participants

Secondary use of large, health-related data sets, particularly those that include genomic data, presents the possibility of inadvertent discovery of a previously unknown health-risk factor that affects the participant. These health risks may be more or less serious, or more or less preventable.

While participant re-identification for return-of-results may be prohibited, or the participant's "right not to know" may need to be enforced, in other ethico-legal contexts, the ability to return results to participants in situations where a serious, preventable condition is discovered may be mandatory. Therein lies the challenge.

A PPRL system thus should be designed to accommodate both possibilities—prohibited re-identification and a required re-identification capability. The most obvious approach would be to design a distributed PPRL system such that the re-identification entity is optional so that, for example, in jurisdictions where re-identification is prohibited, this capability can be omitted. Whatever the approach taken, care must be taken to appropriately account for the potential that a participant has his data held simultaneously by multiple entities, some of which allow re-identification while others prohibit it, to help ensure that the overall system can most optimally comply with the diverging requirements.

3 Technical Methods and Approaches

3.1 Desired Features and Attributes

The joint GA4GH–IRDiRC effort aims to identify and recommend for further consideration one or more approaches to enable linkage of coded[1] data across organizations such that

1. This article uses the term "coded" to describe records or other data whose personal identifiers have been removed and replaced with a re-identification code that is generated independently of the values of identity attributes. Coding requires not only the removal of direct identifiers, but also indirect (or quasi-) identifiers, thereby making it impossible to derive the participant's identity without access to the information associating the code with an individual. Coding, according to this definition, is a form of pseudonymisation.

even though records have been linked, no information about the identity of the individual to whom the data pertain can be ascertained unless the relevant research project has chosen to allow its participants' identities to be disclosed in the limited circumstances where disclosure corresponds to the individual's wishes or where required by law. Consistent with the ethico-legal considerations discussed above, the following desired features and attributes were identified:

- The approach should recognize, with a high degree of confidence, coded records associated with the same individual.
- The approach should be applicable to any data type (e.g., text, clinical data, images, genomic data).
- The approach should use a linkage algorithm that does not require the knowledge of the individual's direct identifiers.
- The approach should not inherently fail to recognize records associated with the same individual due to spelling differences, typographical errors, missing and out-of-date data, and other minor irregularities.
- The approach should enable a participant to limit linkages to her data.
- The approach should use techniques that are resistant to re-identification attacks (e.g., frequency, dictionary, cryptanalysis), while enabling re-identification when required and authorized.
- The approach should enable an assessment of linkage quality and completeness.
- The approach should be scalable and distributable, allowing linkage of very large datasets across multiple organizations.
- The approach should have been implemented for use, and not simply theoretical.

3.2 Current State of Knowledge and Practice

Within research environments like those in which GA4GH and IRDiRC generally work, PPRL is most often used for the purpose of creating a research dataset in which all records pertaining to the same person are linkable, even as the identity of the person remains unknown to the researchers. Accomplishing this presents two different kinds of challenges in an international ecosystem that highly values privacy. The first relates to legal restrictions that sometimes regulate the collection, use, and disclosure of specific attributes including a person's name, gender, birthdate, and place of birth, such as in HIPAA. The second is that the more records a data set contains pertaining to the same individual, the easier it will be to identify that individual—a problem called statistical disclosure control. This tendency simultaneously empowers Big Data analytics.

Experts studying identifiability distinguish between *direct identifiers* such as a person's personal unique identifier (PUID) and, in most contexts, their name, on the one hand, and *quasi-identifiers* (QIDs), such as gender, date of birth, and address, on the other. If a one-to-one mapping of individual-to-code is likely possible, the code is considered a direct identifier. QIDs also, however, play an important role in PPRL.

To be most useful, the records to be linked need to include a common set (or subset) of attributes, and the attributes need to be expressed such that they are recognizable across data sets (e.g., spelling consistency, common metadata, controlled vocabulary). PPRL involving Big Data, such as genomic data, presents additional challenges, including scalability, linkage quality, and increased privacy risk [8]. In some contexts, these challenges can be addressed by pre-processing techniques, such as those described by Christen [9].

Record linkage is generally accomplished using one of three basic types of protocols [9]:

1. Two-party protocols are used when only two database owners want to link their data.
2. Three-party protocols are used when two parties are assisted by a trusted third party, enabling the linkage to occur without either party seeing the other's data.
3. Multi-party protocols are used to link more than two data sets, and may involve a trusted third party.

3.3 PPRL Techniques

PPRL techniques have evolved over time. First-generation techniques (mid-1990s) were primarily based on exact matching using simple hash encoding. The U.S. National Institutes of Health (NIH) Global Unique Identifiers (GUID) approach is an example of this technique [10]. These techniques are challenged by the fact that a single-letter difference in the attribute values used will yield dramatically different hash values.

Second-generation techniques (early 2000s) rely upon approximate matching and include comparisons of edit distances and other string-comparison functions. The principal limitation of these techniques is scalability. Third-generation techniques (mid-2000s) take scalability into account and often represent a compromise between privacy-protection and scalability; these techniques may allow for some information leakage [9].

A large number of matching approaches and protocols have involved some combination or extensions of the following techniques [9]:

- *Secure hash encoding.* This is done using a cryptographic hash function (i.e., a one-way algorithm that, given any size string of characters as input, will produce a unique, repeatable, fixed-size output). Hash functions play a major role in PPRL because of their ability to reliably confirm matching inputs without directly revealing any information about the content of those inputs. However, dictionary and frequency attacks are possible and are generally mitigated by injecting a random value known as a "salt" into the output. Also, hash functions test only exact matches and have no inherent capacity to handle near matches.
- *Statistical linkage key (SLK).* An SLK is a derived variable generated from components of direct and indirect identity elements. The SLK-581 [11], developed by the Australian Institute of Health and Welfare to link health datasets, is an example of a statistical linkage key. The format of the complete SLK-581 is XXXZZDDMMYYYYM—where XXX is the individual's family name, ZZ is the given name, date of birth is represented as DDMMYYYY, and gender is represented as M, F, or U. SLK-based masking has been shown to provide limited privacy protection and

poor sensitivity [9]. Also, because an SLK comprises identity elements, it would not meet the GA4GH requirements that the approach "not require knowledge of the individual's identity" and "use techniques that are resistant to re-identification attacks."

- *Encryption schemes*. These approaches involve the use of encryption algorithms to link data. For example, secure multi-party computation (SMC) is a cryptographic method in which multiple parties jointly compute a function while keeping their individual inputs private. During the computation, each participating party computes part of the function, and in the end, each party knows only the end result and its own input.
- *Bloom filter*. In this approach, hash values are loaded into a vector, which is then compared with other vectors similarly generated, resulting in either a "definite no" or a "perhaps yes" match. The method was first defined by Schnell, whose paper provides a detailed description of how Bloom filters work [1]. Bloom-filter encoding has been widely used as an efficient technique for matching records without sacrificing privacy [1], [9], [10].

3.4 Challenges

Several challenges to accurate and efficient record-linkage have been identified. Van Grootheest et al. [12] studied record-linkage performance under simulated conditions and found that linkage performance is dependent upon the algorithm used, the choice of linkage variables, the dataset size and overlap, and errors in datasets. Although personally identifiable information (PII), such as date of birth and name at birth, often is considered immutable, mistakes and discrepancies in recording the information, language and spelling differences, and variations in format render it mutable nonetheless. Some have turned to biometric data (e.g., fingerprint, DNA) as more immutable for the purpose of linkage. However, this approach can pose its own challenges, including significant privacy risk, collection cost, and inconvenience.

Some approaches are technically interesting, but may be computationally impractical. In particular, approaches involving asymmetric encryption of large quantities of data are computationally intensive. Hardware-based encryption, which implements encryption algorithms in hardware rather than software, dramatically improves the efficiency of encryption solutions; however, software-based solutions are more cost-effective and therefore more widely used. For several decades, homomorphic encryption has held promise for data linking, but has been prohibitively time consuming. Recent advances hold promise for the practical use of homomorphic encryption for data linkage purposes [13].

PPRL approaches are susceptible to several adversarial models and attack methods, such as dictionary attacks, frequency attacks, cryptanalysis attacks, and collusion [9]. As with any system, assessing the security and resilience of any PPRL system may require multiple approaches, depending upon the mechanisms used, the architecture of the system, and the intended use.

Another PPRL challenge is the need to design the system such that linkages can be destroyed or modified once they are established. Such a mechanism might be desirable, for example, to enable new pseudonyms to be reassigned to individuals who were affected by a security breach, to delete a synonym when a participant revokes consent to use her data, or to re-identify and re-contact an individual to provide him with important health-relevant results. This challenge is sometimes addressed by using a hash value as an internal (protected) intermediate value that is associated with an independent, randomly generated identifier that is distributed and used as a pseudonym, with the ability to be reassigned at will. This approach enables a pseudonym to be revoked or changed without affecting every individual in the data set, so long as the hash value itself remains secure.

3.5 Current Approaches

U.S. National Institutes of Health Global Unique Identifier. The US National Institutes of Health (NIH) developed the Global Unique Identifier (GUID) Tool as a customised, client–server software application used to a Global Unique Identifier (GUID) for each study participant. The GUID is a subject pseudonym designed to allow a researcher to share data specific to a study participant without exposing personally identifiable information (PII), and to match participants across labs and research data sets.

To generate a GUID, the data holder enters a number of PII data elements, including gender, name, and birth location, which are used to generate a hash value that the data holder sends to the GUID server. If the hash value matches an existing entry in the GUID index, the associated GUID pseudonym is returned to the data holder. If the hash value does not match an existing entry, a new GUID is randomly generated and returned.

No PII ever leaves the data holder. Because the GUID is randomly generated, no attacker could infer the identity of the individual based on the GUID alone. The same individual's information will produce the same GUID across time, location, and research context, allowing researchers to match shared data regardless of its source without sharing or viewing PII [10]. However, an attacker with knowledge of the data elements used to generate the hash value, and with access to a GUID client, could generate the hash value and then query the server to retrieve the GUID associated with that individual, thus allowing the attacker to re-identify the participant in any pseudonymised data set to which he can obtain access.

Mainzelliste (Germany). Mainzelliste is an open-source, RESTful service for pseudonymisation developed at the Johannes Gutenberg University Mainz. A user inputs PII (e.g., name, date of birth) and receives back a pseudonym generated using data unrelated to the identifiable elements. The pseudonymisation service maintains a database of identifiable data strings matched to pseudonyms. Upon receiving the identifiable elements, the service performs a lookup to determine whether a pseudonym already exists. The lookup runs a linkage algorithm to account for near matches (e.g., typographical errors). If a match is found, the service returns the existing pseudonym. If no match is found, the service generates and returns a new pseudonym [16], [17].

Linkage is possible even in the event of typos or alternate spellings. Mainzelliste allows for the possibility of using in-house phonetic codes and string comparisons for linkage,

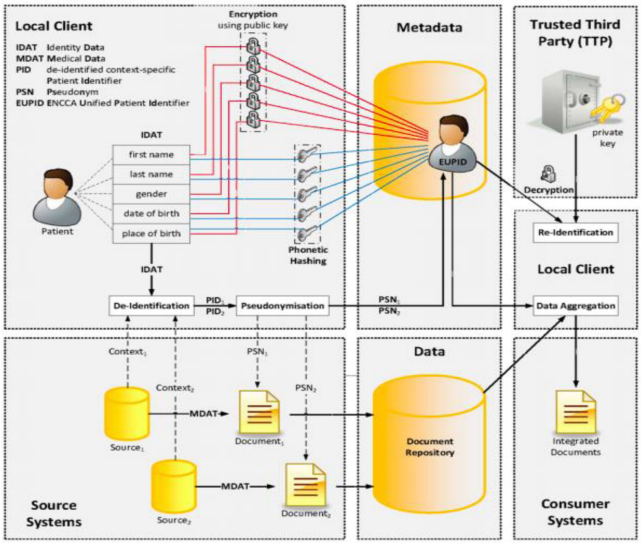

Fig. 1. The ENCCA Unified Patient Identifier (EUPID) approach enables the use of context-specific pseudonyms (PIDs), while preserving the capability for a trusted third party to link PIDs pertaining to the same individual, through the use of indexed EUPIDs [15].

thereby allowing names from other linguistic backgrounds to be fault-tolerantly compared. Currently, weight-based record linkage is supported, but the modular concept allows for retrofitting an in-house algorithm. The possibility to manually rework uncertain assignments further supports the automatic matching process [9].

European Patient Identity Management. The European patient-identity management solution (EUPID) approach addresses the risk associated with the GUID system, as described above, by using context-specific data elements and hashing algorithms to generate a context-specific pseudonym for each individual. The context-specific pseudonyms then are linked within the EUPID system and associated with a linkage pseudonym, without revealing the context-specific pseudonyms included in the association.

The EUPID approach, originally developed by the European Network for Cancer Research in Children and Adolescents (ENCCA), was designed to meet the following requirements:

- Prevent duplicate registration of patients.
- Preserve the capability to re-identify subjects by a trusted third party in special cases.
- Support the capability to use different pseudonyms for the same patient in different contexts, while preserving the capability for a trusted third party to link datasets pertaining to the same patient and stored under different pseudonyms—while assuring that patient identification in any single context is nearly impossible from another context.
- Avoid creating a transparent universal patient ID.
- Assure that the approach can be implemented in a distributed computing environment.

The EUPID scheme is illustrated in Fig. 1, from Nitzlnader and Schreier [15], which provides detail regarding the methodology and how the EUPID linkage is generated and used in actual practice. A key feature is context-specific pseudonymisation, which maintains identity linkages locally, while enabling re-identification of a linked data set through a three-party collaboration involving the local context, the linkage agent, and a trusted third party. EUPID combines several of the PPRL matching techniques discussed in Section 3.3 to identify linkages, and an optional trusted third party to enable re-identification as authorized.

4 Discussion

PPRL techniques and approaches generally use either a direct identifier (PUID) or a quasi-identifier (QID). The use of this information is regulated throughout the world under applicable jurisdictional laws and institutional policies. The mere centralization of data presents privacy challenges. In addition, considerations such as the need to enable participants to withdraw from a research study and the various policies relating to the return of results must be factored into the design of a PPRL approach and implementation.

Any PPRL approach must consider privacy risks inherent in its methods. A hash value generated using PUID or QID cannot be used as a pseudonym because the hash value is derived from personal data, which often is prohibited by applicable law or regulation. Also, use of a hash value as a pseudonym that does not allow re-identification may be illegal in some contexts due to the right of participants to withdraw, to access their own data, and to receive the results of research performed using their data returned when so desired.

Some of the technical approaches examined generate a random or quasi-random pseudonym, and store an association between the generated pseudonym and a hash value derived from PII. In this way, no participant can be identified on the basis of the pseudonym alone. However, a unique linkage between a PII-based hash value and a pseudonym leaves open the possibility of using the linkage system to reverse-discover the pseudonym associated with a known individual.

The Bloom filter approach may be less vulnerable to this type of reverse-discovery attack if a large number of hash functions and a sufficiently long filter are used [9]. The Bloom filter approach produces quantitative values reflecting the strength of the match, and has been shown to produce linkages comparable to those produced by traditional methods using unencrypted identity attributes [12].

The hope behind these initiatives is to simultaneously safeguard privacy while also furthering open data ideals such as the FAIR principles (not to be confused with FIPP, the Fair Information Practice Principles, discussed below), which demand that data be **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable [18]. Since their emergence in the scientific context, FAIR principles have been recognized as particularly important where health-related data are concerned.

It is important to recognize that the very act of linking records pertaining to the same individual may make them easier to identify. Indeed, this is why many research programs require a minimum "bin size" or "cell size"—i.e., that a minimum number of individuals be represented in any bar in a histogram ("bin") or any single "cell" in a spreadsheet. Linking two records within a "bin" essentially reduces the bin size, and increases privacy risk. In addition, a trusted third-party (used in all 3 of the implementations discussed above) becomes an attractive target for attackers.

A strong data-linkage approach should conform to the "principle of least privilege" wherein each entity has access to only those data and system privileges it needs to perform its assigned functions. EUPID takes a step in the right direction through its context-based pseudonymisation and collaborative approach to re-identification. But all three of the approaches discussed in Section 3.5 include a single point of failure. Methods that distribute trust across entities, such as multi-party computation (MPC) and federation, are potential avenues for addressing this vulnerability.

As in other areas of data-sharing, no technical solutions on their own can ensure both data privacy and data sharing. Any workable international PPRL solution will require strong privacy policy, enforceable through the combined use of technical methods, like those discussed above, and robust organizational and governance measures, perhaps taking inspiration from the Fair Information Practice Principles (FIPPs) that undergird most international privacy law.

5 Recommendations

After considering the methods and approaches discussed above, the PPRL Task Force concluded that the EUPID approach held the most promise for the emerging, global GA4GH and IRDiRC research environment. In particular, the Task Force was impressed with EUPID's use of context-specific identity attributes, hashing functions, and pseudonyms to localize privacy risk, and its use of phonetic hashing to enable robust linking. The model in principle can be federated and scaled to accommodate other consortia and data-sharing efforts. In addition, the model's "re-identification" capability could be offered as an optional module for contexts that require the capability to learn the identity of a research participant under special circumstances, and with appropriate authorisation.

The PPRL Task Force is collaborating with the EUPID project to deepen its understanding of the EUPID model and to further explore its use. A security review is planned as well as an investigation of the feasibility of using secure multi-party computation (SMC) as part of the federation model.

Acknowledgments

This work was supported in part by the Can-SHARE project, which is in turn supported by Genome Quebec, Genome Canada, the government of Canada, the Ministère de l'Économie, Innovation et Exportation du Québec, and the Canadian Institutes of Health Research (fund #141210).

References

[1] R. Schnell, "Privacy-preserving data linkage," *Methodological Developments in Data Linkage*, K. Harron, H. Goldstein, and C. Dibben, eds. Hoboken, NJ, USA: Wiley, 2016, pp. 201–225.

[2] Global Alliance for Genomics and Health, "Framework for responsible sharing of genomic and health-related data," 2014, [Online]. Available: https://www.ga4gh.org/ga4ghtoolkit/regulatoryandethics/framework-for-responsible-sharing-genomic-and-health-related-data/.

[3] *Freedom of Information and Protection of Privacy Act*, Revised Statutes of British Columbia 1996, chapter 165, sections 35(1)(b), 36.1(1).

[4] *E-Health Act*, Statutes of British Columbia 2008, chapter 38, section 14(2.1)(*d*).

[5] *Personal Information Protection Act*, Statutes of British Columbia 2003, chapter 63, section 21(1)(c).

[6] *Health Insurance Portability and Accountability Act of 1996 (U.S.)*, Public Law No. 104–191, US Statutes at Large, vol. 110, pp. 1936ff, 1996.

[7] General Data Protection Regulation, *Official Journal of the European Union*, vol. 59, L 119/1. 2016.

[8] D. Vatsalan, Z. Sehili, P. Christen, and E. Rahm, "Privacy-preserving record linkage for big data: Current approaches and research challenges," *Handbook of Big Data Technologies*. Berlin, Germany: Springer, 2016.

[9] P. Christen, "Privacy-preserving record linkage," *ScaDS Leipzig*, 2016, [Online]. Available: http://users.cecs.anu.edu.au/∼christen/publications/christen2016scads.pdf

[10] National Institutes of Health, "Global unique identifier (GUID)," 2016, [Online]. Available: https://data-archive.nimh.nih.gov/guid/

[11] Australian Institute of Health and Welfare, "SLK-581 Guide for use," Australian government, Jul. 2016, [Online]. Available: https://www.aihw.gov.au/getmedia/cf980d57-c72f-4925-b1fc-36be6e80d3cd/aodts-nmds-2017-18-slk-581-guide.pdf.aspx

[12] G. van Grootheest, M. C. H. de Groot, D. J. van der Laan, J. H. Smit, and B. F. M. Bakker, "Record linkage for health studies: Three demonstration projects," 2015, [Online]. Available: http://www.biolink-nl.eu/public/2015_recordlinkageforhealthstudies.pdf

[13] L. Hardesty, "Securing the cloud: A new algorithm solves a major pproblem with homomorphic encryption, which would let web servers process data without decrypting it," *MIT News*, 2013, [Online]. Available: http://news.mit.edu/2013/algorithm-solves-homomorphic-encryption-problem-0610

[14] M. Lablans, A. Borg, F. Ückert, "A RESTful interface to pseudonymization services in modern web applications," *BMC Med. Inf. Decision Making*, vol. 15, no. 2, 2015, doi: 10.1186/s12911-014-0123-5.

[15] M. Nitzlnader, G. Schreier, "Patient identity management for secondary use of biomedical research data in a distributed computing environment," *eHealth2014 – Health Informatics Meets eHealth*, A. Hörbst, et al., eds., Amsterdam, The Netherlands: IOS Press, 2014, doi:10.3233/978-1-61499-397-1-211, [Online]. Available: https://eupid.eu/assets/downloads/nitzlnader2014.pdf

[16] "Mainzelliste," [Online]. Available: https://mainzelliste.de

[17] Institute of Medical Biostatics, Epidemiology, and Informatics (IMBEI), "Mainzelliste as an open source service," University Medical Center of the Johannes Gutenberg University Mainz 2013, [Online]. Archived at: https://web.archive.org/web/20160815024055/http://www.unimedizin-mainz.de:80/imbei/medicalinformatics/ag-verbundforschung/mainzelliste.html?L=1

[18] FORCE11, "Guiding principles for findable, accessible, interoperable and re-usable data publishing version b1.0," [Online]. Available: https://www.force11.org/node/6062

Dixie B. Baker received the PhD degree in special education from the University of Southern California. She is a Senior Partner at Martin, Blanck, and Associates. Her research interests include health information management, digital security, and genomic data sharing.

Bartha M. Knoppers received the PhD degree in law from the Université de Paris I, Panthéon-Sorbonne. She is the director of the Centre of Genomics and Policy at McGill University. Her research interests include governance, genomic medicine, and human rights.

Mark Phillips, received the B.C.L./LL.B. degree in law from McGill University and BS degree in computer science from the University of Manitoba. He is an academic associate at the Centre of Genomics and Policy at McGill University. His research interests include data protection, identifiability, and open data.

David van Enckevort received the Drs. degree in classical archaeology from the University of Leiden. He is a technical project lead at the University Medical Center Groningen. His research interests include biomedical science, system administration, and network administration.

Petra Kaufmann received the doctorate degree in medicine from the University of Bonn. She is the clinical innovation director at the National Institutes of Health NCATS Division of Clinical Information. Her research interests include neuromuscular diseases, clinical research, and clinical trials.

Hanns Lochmuller received the doctorate degree in medicine from the Ludwig-Maximilians-University of Munich. He is a professor at Newcastle University. His research interests include molecular genetics, inherited myopathies, and neuromuscular junction disorders.

Domenica Taruscio received the M.D. degree in medicine from the University of Bologna. She is director of the National Centre for Rare Diseases at the Istituto Superiore di Sanita. Her research interests include histopathology, bioethics, and human genetics.